US010776516B2

(12) **United States Patent**
Dinov et al.

(10) Patent No.: **US 10,776,516 B2**
(45) Date of Patent: **Sep. 15, 2020**

(54) **ELECTRONIC MEDICAL RECORD DATASIFTER**

(71) Applicant: **The Regents of The University of Michigan**, Ann Arbor, MI (US)

(72) Inventors: **Ivaylo Dinov**, Ann Arbor, MI (US); **John Vandervest**, Ann Arbor, MI (US); **Simeone Marino**, Ann Arbor, MI (US)

(73) Assignee: **THE REGENTS OF THE UNIVERSITY OF MICHIGAN**, Ann Arbor, MI (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 227 days.

(21) Appl. No.: **16/051,881**

(22) Filed: **Aug. 1, 2018**

(65) **Prior Publication Data**

US 2019/0042791 A1   Feb. 7, 2019

Related U.S. Application Data

(60) Provisional application No. 62/540,184, filed on Aug. 2, 2017.

(51) **Int. Cl.**
G06F 16/00	(2019.01)
G06F 21/62	(2013.01)
G16H 10/60	(2018.01)
G06N 7/08	(2006.01)
G06F 17/18	(2006.01)
G06F 17/11	(2006.01)
G16H 50/70	(2018.01)

(Continued)

(52) **U.S. Cl.**
CPC *G06F 21/6254* (2013.01); *G06F 16/2272* (2019.01); *G06F 17/11* (2013.01);
(Continued)

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

7,653,615	B2	1/2010	Dwork et al.
7,698,250	B2	4/2010	Dwork et al.

(Continued)

OTHER PUBLICATIONS

X. Yao et al "Differential Privacy of Big Data: An Overview" Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC) and IEEE International Conference on Intelligent Data and Security (2016).

(Continued)

Primary Examiner — Bai D Vu
(74) *Attorney, Agent, or Firm* — Harness, Dickey & Pierce, P.L.C.

(57)   **ABSTRACT**

A method is presented for generating a data set from a database. The method involves iterative data manipulation that stochastically identifies candidate entries from the cases (subjects, participants) and variables (data elements) and subsequently selects, nullifies, and imputes the information. This process heavily relies on statistical multivariate imputation to preserve the joint distributions of the complex structured data archive. At each step, the algorithm generates a complete dataset that in aggregate closely resembles the intrinsic characteristics of the original data set, however, on an individual level the rows of data are substantially altered. This procedure drastically reduces the risk for subject reidentification by stratification, as meta-data for all subjects is repeatedly and lossily encoded.

19 Claims, 3 Drawing Sheets